Speech Recognition System and Isolated Word Recognition based on Hidden Markov Model (HMM) for Hearing Impaired

S.Ananthi
Research Scholar
Dept. of CSE
Annamalai University

P. Dhanalakshmi
Associate Professor
Dept. of CSE
Annamalai University

ABSTRACT

The ability of a reader to recognize written words correctly, virtually and effortlessly is defined as Word Recognition or Isolated Word Recognition. It will recognize each word from their shape. Speech Recognition is the operating system which enablesto convert spoken words to written text which is called as Speech to Text (STT) method. Usual Method used in Speech Recognition (SR) is Neural Network, Hidden Markov Model (HMM) and Dynamic Time Warping (DTW). The widely used technique for Speech Recognition is HMM. Hidden Markov Model assumes that successive acoustic features of a spoken word are state independent. The occurrence of one feature is independent of the occurrence of the others state. Here each single unit of word is considered as state. Based upon the probability of the state it generates possible word sequence for the spoken word. Instead of listening to the speech, the generated sequence of text can be easily viewed. Each word is recognized from their shape. People with hearing impaired can make use of this Speech Recognition.

Keywords:

Automatic Speech Recognition (ASR), Dynamic Time Warping (DTW), Hidden Markov Model (HMM), Information Retrieval, Isolated Word Recognition, Performance, Speech Recognition (SR),Word Recognition.ifx

1. INTRODUCTION

Speech is the vocalized form of human communication and Speech processing is researched in terms of speech production and perception of the sounds which are used in vocal languages. Due to the pressure of the glottis and the air pushed from lungs, the vocal cords can open and close very quickly, which generates vibrations in the air. Speech sounds are analyzed from many points of view namely Articulatory, Acoustics, Phonetic and Perceptual. Human speech production is one of the incredible creatures of god. These speeches can be generated artificially by the computers, called as Speech Synthesis. In contrast to this human speech can be converted into text, called as Speech Recognition. Speech Synthesis and Speech Recognition are incredibly helpful for Impaired Persons. So that, abundance of researches is departure on to make system to speak.

During the precedent existence, Speech Processing and Pattern classification achieves enormous milestones. Plenty of researches were departure on in these pasture and they accomplish sensation in numerous recognition and synthesis techniques and these techniques have collective convention for the people with Speech and Hearing Impaired etc., Subsequently its usage become broadening day by day. Nowadays, impaired persons are using the system because of the great achievement in this territory. Speech recognition is an automated conversion of spoken works into system based readable text format. This technology allows a computer to recognize words that are spoken into microphone or a telephone. This system is also called as Automatic Speech Recognition. Some speech recognition systems recognize only one person's voice which is termed as speaker dependent; others are speaker independent. Large vocabulary speaker independent systems have potential in every form of computing from hand held mobile devices to personal computing and even large scale data centres [12]. Over the past years, Hidden Markov Model (HMM) has been widely applied in several models like Pattern Recognition [16] or speech Recognition [21]. HMM Models a sequence of observations as a piecewise stationary characteristics of word or sub-word units among the various speakers even in large vocabulary.

These systems can be either speaker dependent or speaker independent. If there is individual speaker for training then that process is said to be as Speaker dependent. It recognizes the speech of that trained speakers alone. If the system is trained with multi users, it can recognize the entire speaker's voice. This type of system is called as Speaker Independent [7]. Isolated words recognition is done for a quantity of phoneme. Smallest Meaningful Unit of sound is called Phoneme. Phonemes are trained by the user [6]. Hidden Markov Model is the widely used technique for Speech Recognition. Hidden Markov Models consists of given n number of states which are passing from one state to another, instantaneously at equally spaces time moment. This work is mainly focused for Visual impaired person, so the developers are in a position to widen Speaker Independent system. So,impaired people will benefit from this speech recognition system.

2. RELATED WORKS

The ultimate goal of [24] is to estimate the sufficient statistics to discriminate among various phonetic units while minimizing the computational demands of the classifier. In [24] scaling computations significantly improves the recognition. In [15] Vector Quantization is the technique used to create reference templates reference template for Speaker Recognition. In [17] LPC analysis extracts the features of given words and VQ is used for feature matching in Speech Recognition. By using an improved speech detection algorithm the accuracy of the real time system can be increased significantly. Recognition of Isolated Word using features based on [18] with network classifiers achieves better accuracy than using these features individually. Fisher's Linear discriminant Analysis (FLDA) is a widely used together classification method, [8] [20] is an important method for linear dimension reduction in statistical pattern classification and SR with small and large vocabulary applications. In [22] simple post-processor duration model or a more Complex Hidden Semi-Markov Model based approaches gives the better performance depending upon the efficiency requirements . Hybrid Hidden Markov Model with Conventional DTW achieves the best prototype model during the training phase in order to increase model discrimination [2]. A hybrid end point detector is proposed in [13] which gives less than 0.5 Percent of a rejection rate, while providing recognition accuracy is obtained from hand-edited endpoint. The recognition rate improvement is discussed in [3] by an adaptive technique of combing speech units.

3. DYNAMIC TIME WARPING

Dynamic Time Warping addresses the problem of time alignment, by non-linear one template in an attempt to synchronize similar acoustic segments in the test and reference templates[19]. This DTW procedure combines alignment and distance computation in one dynamic programming procedure. DTW finds an optimal path through a network of possibilities in comparing two multiframe templates. In this small deviation from this linear frame-by-frame comparison are allowed if the distance for a frame pair is smaller than other local frame comparison.

In Speech Recognition system DTW is often used to determine whether the two spoken waveforms represent the same phrase. In a speech waveform, the duration of each spoken sound and the interval between sounds are permitted to vary, but the overall speech has to be similar.

DTW has few advantages for some syllabic utterances but substantial increases in accuracy occurrence for DTW in matching polysyllabic utterances [19].

Most important problem of this scheme is, it can generate only little amount of learning words. The calculating rate of the signal is high and also it requires large memory for storing the speech signal also to generate respective text. DTW are still heavy computational load and treatment of durational variations as noise to be eliminated via time normalization. DTW does not allow weighting different parts of an utterance by their information contribution to Automatic Speech Recognition. Consequently, DTW is an efficient only for some Automatic Speech Recognition.

The user have to construct a HMM which is capable of generating an unlimited sequence of words from the vocabulary or the database. To construct large HMM which reflect the speech recognition task at hand from smaller HMM and performing the recognition task at hand by searching the optimum state sequence for that HMM is one of the most important issues in the stochastic modelling framework.

4. PROPOSED SYSTEM

Hidden Markov Model is proposed in this work. A Hidden Markov Model is considered as a generalization of a mixture model where the hidden variables control the mixture components has to be selected for each observation,these obsrervations are related through a Markov process rather than independent of each other [11]. Hidden Markov Model assumes that successive acoustic features of a spoken word are state independent. The occurrence of one feature is independent of the occurrence of the others [4].

Markov Model is a stochastic model with finite state automaton in which the sequence of states is a Markov chain [14]. Each Markov Model corresponds to a deterministic event, whereas each output of HMM corresponds to probabilistic density function; the generating state sequence of HMM is hidden. Probability starts with a particular event. Markov model is defined as, the states represent possible event types (e.g., the different words in this example) and the transitions represent the probability of one event type following another [5]. It is easy to depict a Markov model as a graph. HMM for Speech Recognition are typically an interconnected group of state. According to an emission probability density function, each state is assumed to emit a new feature vector for each individual frame .

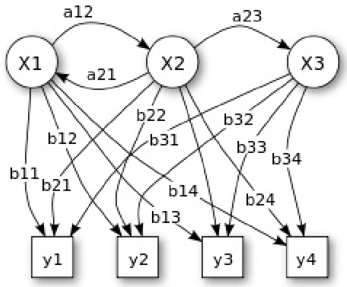

Fig. 1. Probabilistic parameters of a hidden Markov model (example)

Each new observation frame can be associated with any state. However, the topology of the HMM and the associated transition probabilities provide temporal constraints [21]. In this work Speech to text conversion is obtained by Hidden Markov Model (HMM). HMM is widely used technique to convert speech to text. Fig. 1 shows the probabilistic parameters of a Hidden Markov Model. It consist of 3 states namely x1, x2, x3.

In above mentioned Fig. 1, x1, x2, x3 are the states of Hidden Markov Model. y1, y2, y3 and y4 are the possible observations a12, a21, a23 are state transition probabilities and output probabilities are mentioned as b. This is for a general description for 3 states. Consider 4 state Markov model. Four States are: 'Web', 'Mining', 'Data' and 'Based'. Initial probabilities for each states are P ('Web') = 0.7, P ('Data') = 0.3. There are only two initial states. Totaling up of initial state probability should always subsist as 1.

4.1 Isolated Word Recognition

Word Recognition assumes all characters are separated[6]. Consider the word 'Data', this phoneme (smallest Unit of Meaningful word) will consider all characters as separate 'D', 'a', 't', and 'a' likewise. Character recognizer will generate output probability for the occurrence of individual character. There are infinite numbers of observations for an individual character in the pronounced word. Fig.5 and 6 shows the possible observation for the spoken character. In Fig. 2 the probability to obtain 'D' is 0.005. In Fig. 3 the possible observation probability to obtain 'a' is 0.5.

Character recognizer will generate output probability for the occurrence of individual character. There are infinite numbers of observations for an individual character in the pronounced word [10]. Fig.5 and 6 shows the possible observation for the spoken character. In Fig. 2 the probability to obtain 'D' is 0.005. In Fig. 3 the possible observation probability to obtain 'a' is 0.5.
In Isolated Word Recognition, the transition probabilities will be defined differently in two subsequent models. The Observation probability for word is equal to the character recognizer scores [21]. Character recognizer score can be calculated by

$$B = (b_i(\nu_\alpha)) = (P(\nu_\alpha|s_i)) \quad (1)$$

By defining in terms of Lexicon, it construct a single HMM for all words. Hidden states are equal to all characters in the alphabet. From the language model Initial probabilities and Transition

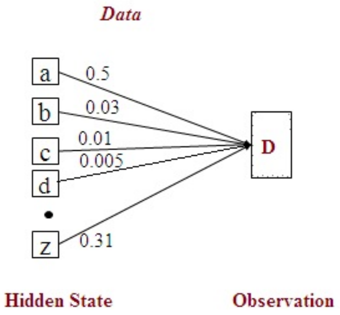

Fig. 2. Observation probability for the word D from 'Data'

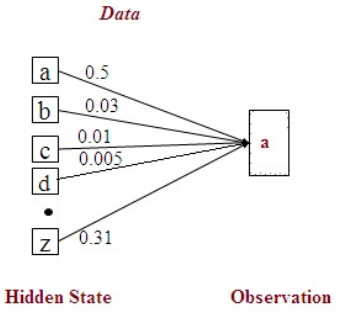

Fig. 3. Observation probability for the word a from 'Data'

probabilities are evaluated. Observations and observation probabilities are as before. Here the user has to determine the best sequence of hidden states, the one that most likely produced word image.

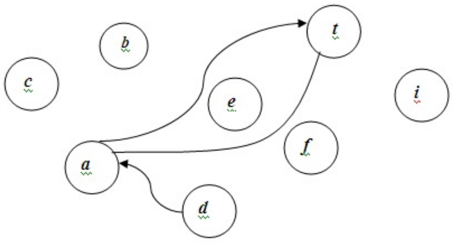

Fig. 4. Isolated word recognition for the word Data

Each and every character in the spoken word is recognized by the system. After the successful completion of word recognition the system combines the huge collection of recognized words together to generate corresponding text for the spoken word.

4.2 Transition and Observation Sequence Probability

In Fig. 1 each rounded circle is a state in Hidden Markov Model. Current state is denoted by Si, next transition state is denoted by Sj. Numbers on arrows between nodes are 'transition' probabilities. Transition probability is calculated by

$$\mathrm{P}(\mathrm{q}_{t+1} = \mathrm{S}_j | \mathrm{q}_t = \mathrm{S}_i) \qquad (2)$$

For e.g., transition probability to reach the state web to mining is $\mathrm{P}(\mathrm{q}_{t+1} = \mathrm{mining}|\mathrm{q}_t = \mathrm{web}) = 0.8.\mathrm{Data}- > \mathrm{Mining}$ is $\mathrm{P}(\mathrm{q}_{t+1} = \mathrm{mining}|\mathrm{q}_t = \mathrm{Data}) = 0.7$.

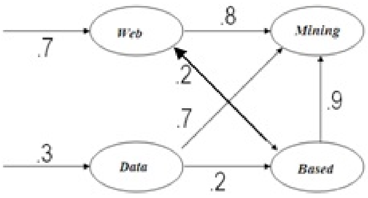

Fig. 5. Example for HMM phoneme (Tiny fragment)

The numbers present in the initial arrows shows the probability of the given state. Missing probabilities are assumed to be 0. The output probability for Web is 0.7 and Data is 0.3. This model has a stochastic transition matrix.

$$A = \begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 0.8 & 0 & 0.2 \\ 2 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0.7 & 0 & 0.2 \\ 4 & 0.2 & 0.9 & 0 & 0 \end{array}$$

1 -> Web
2 -> Mining
3 -> Data
4 -> Based

Fig. 6. Stochastic transition matrix

4.2.1 Computation of the Transition sequence probability. Numbers on arrows between nodes are termed as transition probabilities and these Transition probability can be calculated by $\mathrm{P}(\mathrm{q}_{t+1} = \mathrm{S}_j|\mathrm{q}_t = \mathrm{S}_i))$.

For instance calculate a probability of a sequence of states. i.e., transition state for Data, Based, Web, Mining. Transition Probability for the state P (Data, Based, Web, Mining) is calculated by $\mathrm{P(Mining|Web) + P(Web|Based) + P(Based|Data)}$.

Where,

$$P(Mining|Web) = 0.8$$
$$P(Web|Based) = 0.2$$
$$P(Based|Data) = 0.2$$

Therefore, Probability to accomplish a sequence of state 'Data', 'Based', 'Web', 'Mining' is 1.2 [i.e., 0.8 + 0.2 + 0.2 = 1.2].

4.2.2 Computation of the Observation sequence probability. Observation sequence for state is calculated as AXB. Where A is Initial State, B is Final state or target state and X is the possible transitions between states A to B. For instance calculate the observation sequence probability for Data to Mining. Consider initial state A as Data and the target state B as Mining.

Consider all the possible intermediate state from Data to Mining as X. Therefore Output Probability is $\mathrm{P(Data}X\mathrm{Mining)} = \mathrm{P(Data \leftrightarrow Mining) + P(Data \leftrightarrow Based \leftrightarrow Mining) + P(Data \leftrightarrow Based \leftrightarrow Web \leftrightarrow Mining)}$.

Where,

$\mathrm{P\,(Data \leftrightarrow Mining)} = \mathrm{P}(\mathrm{q}_{t+1} = \mathrm{S}_j|\mathrm{q}_t = \mathrm{S}_i) = \mathrm{P(Mining|Data)} = 0.7$.

$\mathrm{P\,(Data \leftrightarrow Based \leftrightarrow Mining)} = \mathrm{P}\,(\mathrm{q}_{t+2} = \mathrm{S}_j|\,\mathrm{q}_{t+1} = \mathrm{S}_i) + \mathrm{P}(\mathrm{q}_{t+1} = \mathrm{S}_j|\mathrm{q}_t = \mathrm{S}_i) = \mathrm{P\,(Mining|Based)} + \mathrm{P\,(Base|Data)} = 0.9 + 0.2 = 1.1$.

Therefore summing up the above calculated values the obtained Observation sequence probability. P(Data X Mining) = P (Data ↔ Mining) + P (Data ↔ Based ↔ Mining) + P (Data ↔ Based ↔ Web ↔ Mining) = 0.7 + 1.1 + 1.2 = 2.0.

These are the statistical models that output a sequence of symbols or quantities. HMM are used in Speech Recognition because a speech signal can be viewed as a short-time stationary signal or a smallest unit of stationary signals.

HMM is popular technique for Speech Recognition because they can be trained automatically and are simple, computationally feasible to use. Each phoneme will have a different output distribution; a Hidden Markov Model for a sequence of word (phoneme) is made by concatenating the individual trained Markov Model for the separate words and phonemes.

5. DATA SET FOR TRAINING AND TESTING

Speech database is collected for the assessment of the proposed speech recognition system. Speech data's from impaired person was acquired from 20 persons. The data obtained from each of the 20 different speakers is used to train a speech recognition system. The recordings for training and testing the speech recognition models were carried out in separate sessions. Every speech unit is modelled using a HMM. Initially, a set of speech data base are trained. By using this training speech database, the set of HMM is trained. Training intends to create HMMs that model the entire diverse traditions of speech unit's pronunciation for different speakers in different condition.

The set of trained HMMs are evaluated with a testing dataset. These testing issues the recognition rates for the Automatic Speech Recognition system.

6. HMM TRAINING

HMM training procedures are proposed to improve the discriminative power of a HMM without sacrificing its classification capacity. The proposed discriminative HMM consist of a conventially trained model and a discriminative model [9]. In general, the iterative training of HMM emission parameters is significantly dependent on the estimation of the posterior probability.

The training principle is as follows.

(1) Choose a form for the local probability estimator for the densities associated with each state.
(2) Choose an initial set of parameters for the estimators.
(3) Given the parameters estimate the probabilities for each state transition and time. These are essential terms in the estimation of the expectation.
(4) Given the probabilities and the parametric form chosen in step 1, find the parameters. These parameters will be guaranteed to give the best possible improvenment for each model.
(5) Access the new models according to some stopping criterion, if it is not good enough return to step 3.

Although some training approaches use somewhat different criteria and probabilistic estimates, the general form of the training for all statistics, sequence system remains the same.

7. PEOPLE WITH IMPAIRED

People with disabilities can acquire benefit from Speech Recognition for individuals that are Deaf or Hard of hearing (Hearing Disability) [23].

Speech Recognition is used to generate a closed-captioning of conversations such as discussions in conference rooms, class rooms lectures etc., Speech Recognition is used in deaf telephony such as voicemail to text, relay services and captioned telephone [1].

8. EXPERIMENTAL RESULT AND DISCUSSION

The quality of speech recognition is analyzed by its similarity to the human voice and the ability of understanding. The system generates the sequence of text for the spoken word or speech signal. Initially all the words are trained by the system for the conversion of speech to text.

Table 1. Accuracy (in %)for the Spoken Word

Spoken Word	Speaker 1	Speaker 2	Speaker 3	Speaker 4	Speaker 5
Data	88.28	87.63	87.61	86.88	88.03
Mining	87.42	86.81	86.73	87.28	86.18
Web	88.35	86.93	87.76	88.63	86.83
Based	85.86	88.63	87.76	87.19	86.89

After completing the training process system is tested with other users. Each and every words are pronounced by several person and the accuracy is measured based on the pronounced worand the generated sequence of word.

In this work 5 different users(Speakers) were trainned by the Recognition system. After completing both training and testing process system the accuracy is measured based on the spoken word and Recognized word by the system. This system can be used by the person with hearing impaired. In future this work intends to give as input to the web search engine to retrieve the information from the database. After retrieving the required documents the retrieved information has to be spoken, this can be used by the Visual Impaired person.

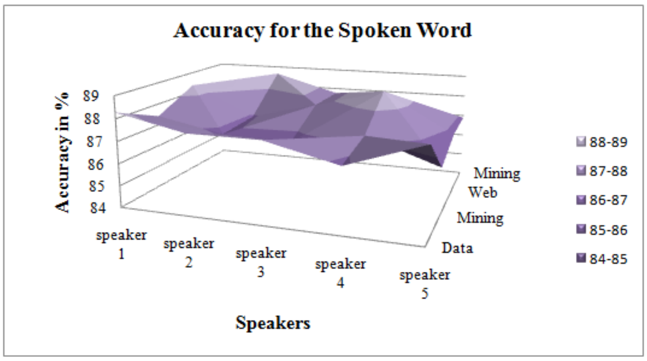

Fig. 7. Accuracy for the Spoken Word from different speakers

9. CONCLUSION

Human and Machine Interactions are concerning the knowledge from environs of Acoustic-Phonetics, Speech Recognition, Speech Perception, Artificial Intelligence etc. The proposed exertion is predominantly focused on Speech Recognition System. The intention of Speech Recognition is to design a system which would be proficient to do continuous speech recognition on huge vocabulary. An attempt has been made through this work for Speech Recognition with large vocabulary. The challenges to the recognition performance of SR are being provided concrete solution so that the gap between recognition capability of machine and human being can be reduced to maximum extend. Speech Recognition based on Hidden Markov Model is achieved successfully for the conversion of Speech to Text. Hearing impaired person can use this Speech Recognition and Speech Synthesis can be very useful for Visual impaired. Speech Recognition is achieved successfully by using Hidden Markov Model. In this proposed exertion Speech Recognition is achieved with 87.42%. In future the generated sequence of word from the human speech can be given as input to the search engine to retrieve information from the web. After retrieving the required documents the retrieved information has to be spoken. This can be utilized by the Visual Impaired person..

10. ACKNOWLEDGMENTS

This work is a part of Major Research Project in University Grand Commission (UGC-MRP 2012). The Author would like to thank Dr. P. Dhanalakshmi who has given a moral support to contribute towards development of this work. The authors would also like to thank the head of the department Dr. V. R. Ramalingam for the Speech and Vision Laboratories who have prearranged high-quality working environment for our Research work.

11. REFERENCES

[1] Language: Implications for deaf readers. *Journal of Deaf Studies and Deaf Education5(1)*, pages 32–50, 2000. Winter.

[2] Hocine Bourouba, Mouldi Bedda, and Rafik Djemili. Isolated word recognition based on hybrid approach dtw/ghmm. *INFORMATICA 30*, pages 373–384, March 2006.

[3] Horia Cucu, Andi Buzo, and Corneliu Burileanu. Optimization methods for large vocabulary, isolated word recognition in romanian language. *U.P.B. Sci. Bull., Series C*, 73(2):179–192, 2011. ISSN:1154-234x.

[4] Baum. L. E and Petrie T. Statistical inference for probabilistic functions of finite state markov chains. *Ann. Mathemat. Stat.*, pages 37: 1554–1563, 1996.

[5] Jelinek. F. Statistical methods for speech recognition. *MIT Press*, 1998. Mass.

[6] Edward Gatt, Joseph Micallef, Paul Micsllef, and Edward Chilton. Phoneme classification in hardware implemented neural networks. *IEEE trans*, page 481, 2001.

[7] Bahlmann. Haasdonk. and Burkhardt. speech and audio recognition. *IEEE trans.*, 11, May 2003.

[8] Heab-Umbach and H. Ney. Linear discriminant analysis for improved large vocabulary continuous speech recognition. *Proc. of International Conference on. Acoustics, Speech and Signal Processing*, 73:13–16, 1992.

[9] Fong Huang and Frank K. Soong. A new discriminative hmm training procedure. *Journal of Acoustic Society of America in 118th Meeting*, pages 481–484, October 1999.

[10] Blimes J. A gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *Technical Report*, April 1998. University of Berkeley.

[11] Rabiner. L. A tutorial on hidden markov model and selected applications in speech recognition. *Proc. of IEEE 37*, page 257.

[12] Rabiner L. and Juang B.H. Fundamentals of speech recognition, prentice-hall. *Englewed cliffs N.J.*, 1993.

[13] Lori F. Lamel, Lawrence R. Rabiner, Aaron E. Rosenberg, and Jay G. Wilpon. An improved endpoint for isolated word recognition. *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSP-29(4):777–785, August 1981.

[14] Baum L.E. An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *in equalities*, 3:1–8, 1972.

[15] Antanas Lipeika, Joana Lipeikiene, and Laimutis Telksnys. Development of isolated word recognition system. *INFORMATICA*, 13(1):37–46, 2002.

[16] Ferrer M.A., Camino J.L, Travieso C.M., and Morales C. Signature classification by hmm. *IEEE International Carnahan Conf. on Security Technology (IEEE ICCST'99)*, pages 481–484, October 1999.

[17] Linga Murthy M.K and Murthy G.L.N. Isolated word recognition using lpc and vector quantization. *International Journal of Research in Engineering and Technology(IJRET)*, pages 479–482, November 2012.

[18] Linga Murthy M.K and Murthy G.L.N. Recognition isolated word using features based on lpc, mfcc, zcr and ste with network classifiers. *International Journal of Modern Engineering Research (IJMER)*, 2(3):854–858, May-June 2012.

[19] Bellman R. and Dreyfus S. Applied dynamic programming. *NJ:Princeton University Press*, 1962.

[20] Raudys S. and Duin P.W. Optimization methods for large vocabulary, isolated word recognition in romanian language. *Pattern Recognition Letters*, 19:385–392, 1998.

[21] Reynals S., Morgan N., Bourland H., and Franco R. Connectionist probability estimators in hmm speech recognition. *IEEE Trans. on Speech and Audio Processing 2(1)*, pages 161–174, 1994.

[22] Levinson S.E. Continuously variable duration hidden markov model for speech analysis. *Proc. ICASSP*, pages 1241–1244, 1986.

[23] Leitch D.and MacMillan T. How students with disabilities respond to speech recognition technology in the university classroom. *Year III Final Research Report on the Liberated Learning Project*, July 2000.

[24] Linde Y., Buzo A., and Gray R.M. An algorithm for vector quantizer design. *IEEE Trans. COM -28*, January 1980.